Title	Validating an academic group tutorial discussion speaking test
Author(s)	Crosthwaite, PR; Boynton, SD; Cole III, SF
Citation	International Journal of English Linguistics, 2016, v. 6 n. 4, p. 12-29
Issued Date	2016
URL	http://hdl.handle.net/10722/226580
Rights	This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

International Journal of English Linguistics; Vol. 6, No. 4; 2016
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Validating an Academic Group Tutorial Discussion Speaking Test

Peter Crosthwaite[1], Simon Boynton[1] & Sam Cole[1]

[1] Centre for Applied English Studies, University of Hong Kong, Hong Kong SAR

Correspondence: Peter Crosthwaite, Centre for Applied English Studies, University of Hong Kong, Hong Kong SAR. E-mail: drprc80@hku.hk

Received: June 3, 2016 Accepted: June 24, 2016 Online Published: July 13, 2016

doi:10.5539/ijel.v6n4p12 URL: http://dx.doi.org/10.5539/ijel.v6n4p12

Abstract

This study attempts to validate an academic group tutorial discussion speaking test for undergraduate freshmen students taking initial EAP training at a university in Hong Kong in terms of task, rater and criterion validity. Three quantitative measures (Cronbach's Alpha, Intraclass Correlation Coefficient, and Exploratory Factor Analysis) are used to assess validity of rater scores for the test using a rubric with considerations for assessment of academic stance presentation, inter-candidate interaction, and individual language proficiency. These results are triangulated with post-hoc interview data from the raters regarding the difficulties they face assessing individual proficiency and group interaction over time. The results suggest that current provisions of the rubric in dealing with the assessment of interaction in group settings (namely visual cues such as "active listening" as well as provisions for interruptions in the form of "domination") are problematic, and that raters are unable to separate the grading of academic stance from the grading of language concerns. We also note affective and cognitive difficulties involved with assessing extended periods of interactional discourse including student talking time (or lack of it), the group dynamic, and raters" personal beliefs and practice as threats to validity that the statistical measurements were unable to capture. A new sample rubric and further suggestions for improving the validity of group tutorial assessments are provided.

Keywords: group tutorial discussion, speaking assessment, English for academic purposes, exploratory factor analysis

1. Introduction

1.1 Use of Group Oral Assessment in Second Language Contexts

There is now a growing trend in the use of peer-to-peer or group oral language proficiency assessments across international and course-specific assessments devised by local teachers (Ducasse & Brown, 2009). Such practice is encouraged from a number of perspectives, including time/cost savings on testing multiple students in one session, the ability of students to engage with Englishes other than the standard variety (Kirkpatrick, 2007), and the avoidance of memorized "interview" talk based on predictable questions (Van Lier, 1989) with unbalanced power relationships between interviewee-interviewer leading to a lack of opportunity for participants to go beyond question-answer conversational structures (Lazeraton, 1992; Johnson, 2001; Taylor, 2001; Galaczi, 2004). From a language acquisition or formative assessment standpoint, group oral assessments allow for the co-construction of knowledge during extended interactional talk, where the comprehensibility of linguistic input (e.g., Krashen, 1987) is enhanced when breakdowns in communication occur via the participant's employment of strategies for negotiation for meaning (Long, 1985) such as clarification requests (*"what you say?"*) or confirmation checks ("High marks?": Pica, 1987). Negotiation for meaning aids students in "noticing" gaps in theirs (or others') linguistic knowledge, which is noted as essential for language acquisition (Schmidt, 1992), and can also lead to the enhancement of student output (Lapkin, 1995), which, in turn, adds to the amount of comprehensible input available during interactional talk. However, Ducasse & Brown (2009) note that the validity of such assessments "depend as much on the criteria and the rating procedure as on the nature of task performance (2009, p. 426), in that while much attention has been paid to discourse analytic studies of peer interaction, little research has yet been conducted on how raters assess such interaction, and the criteria by which such interaction is to be assessed against.

Outside of concerns over the validation of interactional assessments in general, an added element of interest is the assessment of interaction in an academic context, namely that of the group tutorial discussion. In this

academic context, students are expected to formulate a stance or position (e.g., Hyland, 2016) on a given topic, to support their stance with examples drawn from academic sources, to defend their stance in the face of challenges from others in the group, and to critically respond to the stance of others. They are expected to do this using the linguistic expectations of academic discourse, including polite challenges, questions, rebuttals, counter-arguments and presentation of facts, opinions or statistics from academic sources (Hyland, 2005) presented as "spoken citations". Thus, the devices available for interaction in such assessments are highly specialized and limited in scope compared to those of colloquial forms of interaction, and a higher degree of English proficiency is required to be able to manage extended periods of academic discourse. For assessment, these concerns require that raters are aware of what constitutes appropriate presentation, support and defense of stance in academic discourse generally, and that they are sensitive to the management of stance in extended second language (L2) discourse—negotiating the frequent breakdowns in grammar and fluency to be expected of L2 learner production while keeping track on the overarching points the speaker is trying to present. However, attempting to grade for stance, interaction and language issues over extended periods for multiple participants is likely to present a considerable cognitive challenge to the individual rater, and the potential for intra- and inter-rater variability when grading is high. Moreover, the assessment rubric used should be clear in terms of what constitutes "good" performance of stance and interaction when raters make grading decisions, and should also be sufficiently transparent for L2 learners who might use the assessment rubric during in-class peer assessment, or who wish to understand where their own performance can be measured against the expectations of their teachers and course requirements. Issues with rater variability or in negotiating the rubric involved in the assessment would likely result in damage to the overall validity of the assessment exercise.

In this respect, this paper examines the validity of a group oral tutorial discussion assessment currently in use at a leading university in Hong Kong. This is achieved via the triangulation of three quantitative measurements of interrater reliability alongside qualitative interview data taken from the raters themselves, with special reference to difficulties using the criterion-based rubric employed for the test, and of raters" difficulties with the assessment of interaction and of student presentation of stance over extended interactional talk.

1.2 The Core University English Speaking Assessment

The context where the speaking assessment is conducted is via the English language studies centre (henceforth "the Centre") of a leading English-as-a-medium-instruction University in Hong Kong. Prior to 2012, a variety of 3-credit courses were offered by the Centre depending on students" major or need for English, of which a variety of speaking tests (some with and some without rubrics) were in place. In 2012, the university began to offer a mandatory "Common Core" curriculum of key academic and social subjects from students" sophomore year that run alongside students" selected major subjects. This new curriculum required a greater involvement with academic English by students of all disciplines. As a response to this need, a new mandatory semester long 6-credit general EAP course of 36 contact and 120 learning hours ("Core University English")was introduced to all freshman students, approximately 2800 students over the two semesters of the freshman year (approx. 1400 students per semester). The course contains three main assessed components, namely, a mid-term written essay or report worth 20% of the total grade, a final written essay or report worth 40% of the total grade, and a group tutorial discussion speaking test worth 40% of the total grade. The speaking component of the course involves 5 taught units (via textbook and in-class activities) involving analyzing the speaking test rubric (unit 1), analyzing the difference between written and spoken academic language and revising written note-taking into spoken form (unit 2), identifying and producing criticism and challenges to others" stance (unit 3), linking your production to what others have said and practicing critical questioning (unit 4) and reviewing / reflecting on discussion skills (unit 5).

The main spoken component of the course is that of the group tutorial discussion. The decision to adopt this kind of activity was taken by the course developers so that it would reflect the typical speaking scenario students would be expected to encounter during their Common Core curriculum following their freshman year. The tutorial discussions ask students to do some background reading on a supplied topic 72 hours prior to the actual discussion, making bullet-point notes and citing their academic sources on a single A4 page-sized template that students bring to the discussion. Students then sit together to discuss this topic in groups of five (although this occasionally drops to four a student is sick/absent). Students perform this type of discussion on different topics at a number of points throughout the course. During Units 1-3, they hold a discussion for 15 minutes, then a mock speaking test is performed in week 10 (after the taught components) under exam conditions of 25 minutes, and the actual assessed speaking test of 25 minutes is held the following week. The physical arrangement setup is as shown in the figure below:

Figure 1. Group tutorial discussion arrangement

For the actual speaking assessment, the discussions are video-recorded and teachers are asked to watch the full video and to enter the marks for each student into an online portal. Individual student performance is graded, rather than the group's production as a whole. Teachers always rate students that were taught by other colleagues. Teachers may watch the videos more than once, but due to time constraints, the majority of teachers watch each video only once.

1.3 The Group Speaking Assessment Rubric

The Centre uses rubrics for both the writing and speaking assessments, although the present study is focused on the speaking assessment rubric only. The rubric is included in Appendix A. The rubric was created collaboratively by course teachers during the creation of the CUE course, and was the first time that such a rubric had been developed by staff at the Centre. The rubric uses a combination of holistic and analytic approaches. The rubric is analytic in the sense that it currently consists of three main scales that each contribute towards the final grade, although the weighting is not equal between all three. These are labeled *ability to explain academic concepts and stance* (worth 40% of the total), *ability to interact with others* (30%) and *ability to communicate comprehensibly and fluently* (30%). The original version of the rubric had an additional scale (*Ability to use academic courses to support a position*) but use of sources was incorporated into the first scale (*ability to explain academic concepts and stance*) from the 2015/2016 academic year. The selection of these scales was done in consultations between course creators based on considerations of good practice in EAP, rather than being modeled on an existing standardized, validated assessment framework. Grades for each scale follow a letter score from A+ (highest) to D- (lowest) or F (fail). Each scale has a number of sub-criteria that need to be taken into account when determining the grade (A+ to F) for that scale. For example, three of the sub-criteria for the scale *ability to communicate comprehensibly and fluently* include "You are always comprehensible", "Mistakes with grammar / vocabulary are infrequent and never interfere with understanding", and "You are always fluent". The exact wordings of these changes depend on the grade to be awarded, with A+ to A- grades require that students are "always" comprehensible, for example, B+ to B- grades "nearly always", etc. However, separate grades are not awarded for each sub-criteria, and a holistic grade is instead awarded based on the combined contribution of each sub-criteria to the overall scale. The intention of the course creators was that the selection and wording of the sub-criteria should aid teachers in the holistic interpretation of the larger super-scale. This was considered particularly important in 2012 as 11 new teachers were hired to teach the course due to the massive increase in workload required. More importantly, however, the rationale behind the rubric and the selection and wording of the sub-criteria is that they are used in student analysis, discussion and criticism of what constitutes best practice in spoken academic group tutorial discussions, allowing students to assess both their own performance and that others" performance via self- and peer-feedback using the rubric as a guide, and promoting assessment of and for learning.

However, despite the assumption that the use of a rubric will serve to increase rater reliability, both teacher and student raters often struggled to reach consensus on an overall impression. It was felt by many that the criteria was open to subjective or ambiguous interpretations. Rater variability has sometimes been shown to increase after implementing a rubric (Knoch, 2009; Wiegle, 2002; Rezaei & Lovorn, 2010). This is also the case where sub-criteria are to be used to reach a holistic overall decision, with raters interpreting the weighting of criteria

differently despite similar training or language background (Brindley, 1991; Hamp-Lyons, 1989; Cumming, 1990; Davison, 2004), misinterpreting the language of the criteria (Lumley, 2002; Kohn, 2006), making judgements on the criteria based on personal beliefs and teaching experience (Davison, 1999, 2004; Arkoudis & O'Loughlin, 2004) or with students performing very well under some of the sub-criteria but poorly in others (Saxton, Belanger, & Becker, 2012). Previous informal correspondence with teacher raters suggested that each of the above issues was a factor when considering the validity of the test as a whole.

1.4 Rationale for Present Study

Prior to the implementation of the rubric in 2012, no trialing had taken place. Rather, following implementation, moderation activities are held with all CUE teachers (*n*=30-35, depending on student numbers/workload) each semester. Moderation involves viewing a single video from a past semester and grading each student, then meeting to discuss the variation against the grading decisions of the course coordinator and a small organizational team. However, previous moderation meetings often resulted in large (and vocal) disagreements among staff about the grading for individual students and of interpretation of sub-criteria generally or in particular instances, with many staff members as much as one full letter grade above or below the moderation team's "model" interpretation of individual students' performance on each criteria. Teachers were not asked to re-do the moderation if they were over/under the average. In the real test, different teachers do not grade the same students, and so there is no current data available on intra- or inter-rater reliability on grading decisions generally, nor has any validation study been carried out on the rubric as a whole, the three main scales, their sub-criteria, or the individual grade descriptors. With this in mind, the present study seeks to answer the following research questions:

1) What level of consistency, consensus and measurement agreement in terms of intra- and inter-rater variance is found as a result of the CUE group tutorial discussion, and how can we account for any significant variance?

2) Does the criterion validity of the assessment rubric prevent raters from accurately and fairly assessing student performance over an extended period of interactional talk? If so, which criteria are considered problematic?

3) How can the overall validity of the assessment be improved with relation to rater reliability and criterion validity?

2. Methods

2.1 Assessment Population of Interest

The present study uses two full 25 minute speaking assessments where the performance of 10 students (5 per video) is to be determined. This is a small proportion of the overall 1400 students that take this assessment each semester, yet given that the rating exercise for two videos would take over an hour of the raters" time, we decided that two videos was the most we could ask volunteer raters to grade. The authors of the present study contacted their former students by e-mail to ask if they would give their consent for the recordings of their real speaking tests to be used for research purposes. At the time of the recordings, these students were all freshman in their first semester of study at the university, a mix of men (n=6) and women (n=4).

2.2 Rater Population of Interest

14 raters from the population of 30 who currently teach the course in question volunteered to rate the two videos and to join in the post-hoc interviews. Table 1 shows the characteristics of the rater sample provided by raters in the questionnaire.

Table 1. Rater sample characteristics

Characteristics	Frequency	% of sample
Age		
20-30	2	15.4%
30-40	5	38.5%
40-50	5	38.5%
50-60	1	7.7%
Gender		
Male	9	69.2%
Female	4	30.8%
L1 Background		
English	7	53.8%
Cantonese	5	38.5%
Other	1	7.7%
Opinion of native speakerness		
Native	8	61.5%
Non-Native	5	38.5%
If N-N, length of English proficiency		
25-30	2	40%
30-35	2	40%
35+	1	20%
Years of English teaching experience		
<5	2	15.4%
5-10	2	15.4%
10-15	5	38.5%
15-20	4	30.8%
Experience of professional language rating		
Yes	9	30.8%
No	4	69.2%
Years of experience teaching course.		
<1 year	2	15.4%
2 years	3	23.1%
3 years	5	38.5%
4+ years	3	23.1%

2.3 Measurement Process

Raters were invited individually or in pairs into a room with a large screen TV where two full video-recorded assessments were played. Raters were given an A3-sized note sheet containing gaps to enter grades for each student across each criteria. Under real assessment conditions, raters are asked to holistically grade the three main criteria (Stance, Interaction, Language) using the (unweighted) sub-criteria as a guide when making their grading decisions. For the purposes of the validation study, however, raters were invited to offer grades for each of the sub-criteria individually, without omitting a grade for any of the sub-criteria. Raters could make notes as they watched the video on a separate note sheet. Immediately after the first video was complete, raters were invited to assign their grades if they had not already done so. They were then asked a series of interview questions by a research assistant designed to gauge their opinions on the assessment and grading process. They were then invited to watch and grade the second video under the same conditions, and were interviewed again immediately following this. The grades and interview data were transcribed by a research assistant, and were coded under the following coding scheme (Table 2) by the first author and the research assistant.

Table 2. Coding scheme for interview data

Coding Scheme	Gloss	Example
Reference to assessment / rater		
Change of mind / grades	Rater changes their mind about a previous grading decision	*I gave him a B- for his language. I think I'll change it to B because he's not as bad as I thought.*
Difficulty of experiment	Rater comments on difficulty with the think-aloud task itself	*In real life I wouldn't be verbalizing my thoughts, so I wouldn't be obscuring what they're saying*
Difficulty of criterion interpretation	Rater comments on difficulty interpreting individual criterion	*[Never read from your notes] Its such a hard thing to grade that one because it's a visual thing, same with listening skills*
Affective difficulties	Rater comments on cognitive overload	*I think its just a case that you are listening to so many things at the same time* *She's pretty good so far*
General appraisal of performance	Rater appraises student performance in general terms without explicit reference to criteria	*I'll be more positive about students if I hear things that have come out on the course*
Comment on personal beliefs	Rater comments on the personal beliefs or attitudes of the teacher / rater	*When the guy starts talking, I'm picking up on whether he's linking what it is that he saying*
Reference to rater practice/Reference to grading decision	Rater comments about the process of rating or about actual grading decision	*I thought at times he was struggling a bit for words I"ll go B minus there instead.*
Reference to criteria		
Ability to express academic concepts and argue for a stance supported with sources	Rater comments specifically on main or sub-criteria related to this rubric section	*The point she is making are not really supported by general world knowledge*
Ability to interact with others	Rater comments specifically on main or sub-criteria related to this rubric section	*I definitely like the way he's tried to link this idea with what came before*
Ability to communicate comprehensibly and fluently	Rater comments specifically on main or sub-criteria related to this rubric section	*I think that maybe language made it a little bit tricky at times to fully follow what he was saying*

Two other native English speaking colleagues were recruited to check the accuracy of coding for each comment (either correct / incorrect) across the entire data, with an Intraclass Correlation Coefficient of .825 suggesting a high level of agreement on the accuracy of the coding.

2.4 Statistical Analysis Adopted (ICC, Cronbach's, EFA, Questionnaire, Interviews)

The present study adopts three statistical measures of interrater reliability on raters' grade assignments according to the criteria on the assessment rubric. The choice to use a variety of statistical measures is inspired by Stemler (2004), who suggests that triangulating consensus, consistency, and measurement approaches to interrater reliability is preferred to using a single unified measure. Consensus relates to raters" ability to agree on which level of a given criteria a given performance should be awarded (or rater bias, in other words), while consistency relates to a given raters' ability to provide consistent grades for performances of the same level, and measurement estimates relate to how raters' decisions across all criteria contribute to the final grade awarded. The triangulation of these three measures of interrater reliability thus accounts for variance in terms of the raters' views of student performance, variance in terms of interpretation of individual criteria, and a combined measure of variance in respect to both performance and criteria.

For the calculation of consensus, the present study adopts the Intraclass Correlation Coefficient (ICC). Stemler (2004) outlines the advantages of using percentage agreement or Cohen's Kappa (Cohen, 1960) to measure consensus, yet Landers (2015) suggests that use of Cohen's Kappa requires scores from two raters only, while we wish to determine consensus over 14 raters. Moreover, the use of Cohen's Kappa is not ideal for interval or ratio scale grades such as the letter grades assigned in the speaking assessment validated in the current paper. To validate the test rubric, we calculate the ICC on each of the individual main and sub-criteria found on the rubric (n=10) in order to determine which criteria lacks rater consensus across the students' performance as a group. This procedure follows Landers (2015) in adopting a two-way random ICC measure as the raters were consistent for all rates and criteria, and adopting the "absolute agreement" variable over the "consistency" variable for ICC

measures in order to capture the consensus on each criterion as an individual factor rather than across a collection of criteria.

For calculation of consistency, we adopt the use of Cronbach's Alpha (Crocker & Algina, 1986). This is used to measure how the ratings of a group of raters serve to determine the reliability of grading decisions across the whole set of main and sub-criteria found on the rubric. Given that in the real test, scores for participants are holistically summarised across the main criteria only, Cronbach's Alpha measures whether the raters were at least consistent in achieving this goal (intra-rater reliability), and if a high Cronbach's Alpha score is achieved for a criterion (i.e., a Cronbach's Alpha value of >.700), then the criterion can be said to be valid. This measure requires each rater to provide a score for each student under each criteria, which was the approach adopted in the present study. ICC is claimed in Stemler & Tsai (2008) to confound both consensus and consistency. However, ICC will be used in the present study to determine consensus of rater grades across individual criteria only, as the ICC value across all criteria is equal to that of Cronbach's Alpha, and, unlike ICC, the calculation of Cronbach's Alpha in SPSS allows one to determine how much the Alpha value would be increased if a particular criterion was removed from the whole set.

For calculation of measurement estimates, the present study adopts an exploratory factor analysis (EFA). This approach seeks to determine the shared level of variance in the ratings that can be accounted for by statistical theoretical latent factors computed from the dataset as a whole (including discrepant ratings) so that one can determine the underlying construct of interest. The variance explained by these latent factors indicates where agreement on the construct of interest is found, and high factor loadings of individual criteria within these factors indicate where agreement on the criteria is grouped. Given that there are three main criteria on the rubric, one would expect an EFA to account for these three criteria as latent factors, and the EFA would determines whether agreement on these three main criteria on the rubric is reached (and does not carry over into other criteria, given that each main criterion is awarded a particular percentage of the overall grade). The EFA would also allow us to consider whether (dis)agreement on individual sub-criteria is sourced under the umbrella of their respective main criteria, or if sub-criteria are being considered in light of another main criteria, if it all. Landers (2004) encourages the use of the many-facets Rasch model (Linacre, 1994) which includes measurement of factors (or "facets") including examinees, tasks, interviewers, scoring criteria and raters to provide an overall picture of what is measured (Eckes, 2009). Rasch Analysis has already been conducted on a group speaking activity designed for Japanese students of English (Bonk & Ockey, 2003; Van Moere, 2006) using a large sample size of raters and examinees. However, Linacre (2004) suggests a minimum sample size of 30-50 raters/examinees for this procedure to be effective, which was unobtainable for the present study. Moreover, Sawaki (2007) suggests that unidimensional models (such as many-facet Rash measurement) "do not seem to be fruitful […] what one is communicating loud and clear by employing such scales is the presence of more than one ability of interest." (2007, p. 359). Given these concerns, EFA was the preferred method for measurement analysis.

3. Results

3.1 Descriptives

Grades for each criteria follow a letter scale from A+ (highest) to D- (lowest). No F grades were reported for any student under any criteria. These grades were then converted into a number with D- = 1, D = 2, D+ = 3, A- = 10, A = 11, A+ = 12.

In order to determine whether the average mean ratings for each criteria in Video 1 were significantly different from those of Video 2, an ANOVA test was employed across the 10 criteria, using a corrected alpha value of .005 for significance due to considerations for multiple testing. No significant differences were reported between the average ratings for videos 1 and 2 for any of the ten criteria, and so the results for both videos are reported together in Table 3 and for the following validation analyses unless clearly specified in the text.

Table 3. Descriptive statistics for ratings across each criteria

Criteria	Mean/ SD
Stance	
Explain academic concepts	9.4 (1.24)
Argue for stance with sources	9.5 (1.35)
Critically respond to others" stance	9.3 (1.48)
Interaction	
Never dominate the discussion	10.7 (1.55)
Never read from your notes	10 (1.42)
Link contributions to what has been said before	9.8 (1.38)
Use active listening skills	10 (1.30)
Language	
Always comprehensible	9.1 (1.37)
Mistakes of grammar/vocab	8.9 (1.07)
Fluency	8.9 (1.31)

The average grade for the *Stance* criteria is 9.4 —a letter grade of B+ —while *Interaction* averages at 10.1 (A-) and *Language* averages 8.9 (B). However, as the weightings for *Stance* (40% of the total grade) are higher than those of *interaction* and *language* (30% each), the average letter grade for the sample across the three criteria is B+.

3.2 Validation Results—Quantitative Approaches—Intraclass Correlation Coefficient

To determine the consensus of ratings on each individual criteria on the assessment rubric, separate ICC analyses were performed on raters" rating of each student in both videos for each criteria (Table 4). The "average measures" value is the one reported Table 4 as we analyse mean data from a sample of the raters and not the whole population of raters. Also, we include the ANOVA F Test values used to determine the random effect of raters and scores for each analysis. The ICC measure itself translates into how much of the variance between raters is "real" in terms of the ratings assigned to it, with higher values demonstrating increased internal consistency while low values demonstrate low internal consistency between raters (A score of below .500 may then be considered as less than chance). Reliability coefficient scores of >.600 are used as a minimum standard by some (Tsai, 2008; Kotner, Audige, Brorson, Donner, Gajewski, Hrobjartsson, Roberts, & Streiner, 2011) while coefficient scores of between .600 and .740 are considered "good", while scores of > .740 are considered "excellent" by others (Fleiss, 1981).

Table 4. Intraclass correlation coefficient per criterion

Criteria	Intraclass Correlation Coefficient	ANOVA F Test values
Stance		
Explain academic concepts	.654	$F(9, 108)=3.204$, $p=.002$
Argue for stance with sources	.726	$F(9, 108)=4.050$, $p<.001$
Critically respond to others" stance	.791	$F(9, 108)=5.163$, $p<.001$
Interaction		
Never dominate the discussion	.453	$F(9, 108)=2.519$, $p=.012$
Never read from your notes	.362	$F(9, 108)=1.955$, $p=.052$
Link contributions to what has been said before	.732	$F(9, 108)=4.461$, $p<.001$
Use active listening skills	.688	$F(9, 108)=3.730$, $p<.001$
Language		
Always comprehensible	.762	$F(9, 108)=5.222$, $p<.001$
Mistakes of grammar/vocab	.807	$F(9, 108)=5.814$, $p<.001$
Fluency	.876	$F(9, 108)=11.407$, $p<.001$

The ICC results suggest that the internal consistency of ratings for each individual criteria are high, except for the criteria *Never dominate the discussion* and *Never read from your notes*, with ICC measures of <.600. In fact, an insignificant F test is noticeable for the criterion *never read from your notes,* suggesting that raters do not demonstrate a significant level of consistency when rating for this criterion.

3.3 Cronbach's Alpha

Cronbach's Alpha analysis was performed on all criteria, firstly across both videos, then for each video

individually. Over both videos, a total Cronbach's Alpha value of .817 was achieved for the 10 criteria, which is considered "good" under George & Mallery's (2003) rules of thumb (Note 1). Table 5 below describes the descriptive data of the Cronbach's Alpha analysis:

Table 5. Cronbach's Alpha scores across criteria

Criteria	Scale mean if item deleted	Scale variance if item deleted	Cronbach's Alpha if item deleted
Stance			
Explain academic concepts	86.69	55.26	.780
Argue for stance with sources	86.56	53.72	.776
Critically respond to others' stance	86.70	53.24	.781
Interaction			
Never dominate the discussion	85.33	63.20	.839
Never read from your notes	86.00	61.27	.823
Link contributions to what has been said before	86.29	55.79	.791
Use active listening skills	86.00	59.39	.807
Language			
Always comprehensible	86.91	59.24	.810
Mistakes of grammar/vocab	87.13	58.73	.793
Fluency	87.19	57.81	.799

Of interest, there are two criterion that, if deleted, would raise the Cronbach's Alpha value, resulting in a statistically higher internal consistency. Namely, *Never dominating the discussion* and *Never reading from your notes* may be safely removed from the rubric, and doing so would improve the Cronbach's Alpha to .855 from .817. The Cronbach's Alpha analysis was then conducted on the results for each individual video. For video 1, a Cronbach's Alpha of .819 ("good") was achieved, and for video 2, a value of .803 ("good"). For video 1, removing the criteria *dominating the discussion* would result in a higher Alpha value of .840, and for video 2, removing the criteria *dominating the discussion* and *never reading from your notes while expressing your stance* would result in a higher alpha value of .875. These results suggest that our raters do not grade these criteria as consistently as they do other criteria.

However, according to Gliem & Gliem (2003), a high value for Cronbach's alpha "indicates good internal consistency of the items in the scale, [yet] it does not mean that the scale is unidimensional" (p. 86). Factor analysis is thus required to control for the dimensionality of a scale along the three perceived dimensions (*Stance, Interaction, Language*) found on the assessment rubric.

3.4 Exploratory Factor Analysis

Table 6 displays the factor loadings for each of the criteria as a pattern matrix following the EFA. The EFA used the Maximum Likelihood extraction method based on factor Eigenvalues greater than 1 for inclusion in the subsequent pattern matrix. Three factors reached Eigenvalues of >1. As correlation between factors were likely, a Promax rotation (Kappa=4) was preferred in this EFA, taking 5 iterations to complete. Factor loadings of .40 were considered a sufficient for the inclusion of a criteria on a given factor (loadings of <.40 are represented by the blank spaces in the pattern matrix below). The Kaiser-Meyer-Olkin Measure for Sampling Adequacy gave a KMO value of .826 (excellent), and Bartlett's Test for Sphericity gave an approx. Chi Square = 588.714, df. 45, p<.001, making the data sufficient for exploratory factor analysis. These factors are cumulatively responsible for 71.281% of the variance explained (Factor 1 = 42.145%, Factor 2 = 16.624%, Factor 3 = 12.512%), which is considered acceptable in Stemler & Tsai (2008).

Table 6. EFA results for speaking test criteria (Videos 1 and 2)

Criteria	Factor 1	Factor 2	Factor 3
Stance			
Ability to critically respond to others" stance	.874		
Ability to link contributions to what has been said before	.867		
Ability to argue for stance with sources	.752		
Ability to explain academic concepts	.575	.459	
Use active listening skills	.505		
Always comprehensible		.847	
Fluency		.807	
Mistakes of grammar/vocab		.660	
Never read from your notes			.761
Never dominate the discussion			

The results of the EFA suggest that there is significant communality between the three criteria for stance (*ability to explain academic concepts, ability to respond to other's stance*, and *using sources to support a stance*), and two of the criteria for interaction (*ability to link turns* and *active listening*). There are a number of implications of this result. Firstly, as mentioned above, Factor 1 is responsible for 42.145% of the total variance explained in the EFA, suggesting that the grading of stance / interaction is problematic for the raters involved when compared to the grading of language issues (Factor 2), which only account for 16.624% of the total variance explained. Secondly, our raters may be double dipping when making grading decisions for the criteria of stance and interaction, in that the communalities for each (with the exception of *taking notes* and *dominating the discussion*) are sufficiently high along factor 1 so as make these (initially separate) criteria now indistinguishable, at least according to the EFA.

Another interesting finding of the EFA is that the criteria *Ability to explain academic concepts* also has a high loading on Factor 2 (accounting for 16.624% of the total variance explained), which is made up of the criteria for language issues. In this respect, while it is perhaps unsurprising that high marks for language issues are correlated with successful explanation of academic concepts, this result suggests that raters are again double dipping when it comes to grading the criteria of stance and grammar.

Another feature highlighted in the EFA is that of factor 3 (accounting for 12.512% of the total variance explained), which has a high loading for the criterion *never reading from your notes while expressing your stance*, but not for any of the other interaction criterion. This again suggests that while the criteria for interaction are treated as a separate consideration to stance and language issues in the marking scheme, in actuality, the raters only consider reading from notes as separate to the criteria for stance. In addition, the results for the criterion *never dominating the discussion* do not result in a significant loading for the criteria on any of the factors produced by the EFA. This suggests that this criterion can be removed from the rubric with no loss of validity for the assessment criteria, either for interaction specifically, or as a whole.

3.5 Explanation of Variance-Rater Characteristics

The quantitative findings suggest issues with consistency, consensus and measurement across the criteria. In attempting to provide qualitative explanations for the quantitative findings, the first approach was to determine whether any of the rater characteristics sampled in the initial questionnaire were predictors of variance (in the form of significantly higher/lower grade averages) for each criterion. An alpha value of 0.005 was used for significance (tests=10, 1 for each criterion) for each variable. As these statistics were not in a normal distribution, the significant values for each Kruskal-Wallis test are provided in Table 7, with p values following Holm-Bonferonni correction for multiple pairwise comparisons:

Table 7. Effect of rater variables on scores awarded per criterion

Characteristics	Criterion affected	Kruskal-Wallis	Pairwise Comparison
Age	*Never dominate the discussion*	H(3)=18.737, $p<.001$	30-40<20-30
Gender	No criterion affected		t(6)=20.65, z=3.877 p=.006
L1 Background	No criterion affected		
Opinion of native speakerness	No criterion affected		
Years of English teaching experience	*Never dominate the discussion*	H(3)=20.739, $p<.001$	10-15<less than 5
Experience of professional language rating	No criterion affected		t(6)=26.50, Z=3.877, p=.006
Years of experience teaching course.	No criterion affected		

From these results, it appears as though the configuration of rater variables does very little to affect the grades of individual criteria, with the exception of *never dominate the discussion* and only for the variable of age (with 20-30 year olds rating higher for this criterion than 30-40 year olds) and years of teaching experience (with those who have less than 5 years of experience rating higher for this criterion than those with 10-15 years of experience). We assume that older and more experienced teachers do not consider domination to be an important feature of the rubric compared with younger and less experienced teachers who may be more likely to more strictly interpret certain criteria as a result of standardisation. However, of more importance, there is no statistical correspondence between other rater variables and criterion scores. With this in mind, we now turn to the qualitative interview data to determine the potential source of the variance found.

3.6 Rater Interviews

Table 8 describes the frequency of coded comments from raters from their post-rating interviews:

Table 8. Descriptive statistics of coded segments for interview data

Comment code	Frequency (Note 2)	Percentage per section of scheme
Reference to assessment/rater		
Change of mind / grades	8	1.9%
Difficulty of experiment	21	4.9%
Difficulty of criterion interpretation	75	17.5%
Affective difficulties	71	16.5%
General appraisal of performance	39	9.1%
Comment on personal beliefs	86	20.0%
Reference to rater practice	63	14.7%
Reference to grading decision	66	15.4%
Reference to rubric criteria Stance		
Ability to express academic concepts	25	10.5%
Ability to argue for a stance supported with sources	34	14.2%
Ability to respond to others" stance	18	7.6%
Interaction		
Never dominate the discussion	25	10.6%
Never read from notes	7	2.9%
Always link contributions to what was said before	24	10.1%
Always use active listening skills	22	9.2%
Language		
Always comprehensible	41	17.2%
Mistakes with grammar / vocabulary	12	5.1%
Always fluent	30	12.6%

3.7 Reference to Assessment

In total, there were 429 coded comments regarding the assessment, assessment practice or the rater. Key themes that arose from this section of the data were related to the time needed to grade the criteria, the correlation between grading and student speaking/action time, and issues with the group dynamic.

In terms of the time needed to grade the criteria, raters commented that more time was needed if they were to assign grades properly, and that they would need to re-watch the performance to do this:

> (R5) *If I am required to give a grade for each major criteria like "*ability to explain academic concepts*",* "bility to interact with others", "ability to communicate comprehensively" *then I don't need to watch the video again, but then if I have to go to each of the smaller criteria then I may need to watch the video more than once.*

But watching the video again was not considered to be a useful exercise by all raters:

> *(R2) I won't [watch it again], I think it is more reliable if I just give a grade right after the discussion. If I revisit the video, if I grade them again, I may struggle between which criteria is better than the other, so I trust my own judgement for the first instance.*
>
> *(R13) but the point is we shouldn't be watching the video twice, [we] should have a test that allows us come to a right conclusion after one view.*

Ideally, given the large numbers of tests conducted each semester and the time taken to grade each one, the

assessment exercise should only be conducted *once* per video, or the exercise requires too much manpower. If raters feel they are forced to re-watch the videos multiple times to arrive at a "safe" grade, then the test cannot really be considered reliable, given that raters may change their grades across multiple viewings as suggested by rater 2 above.

Another large factor in grading difficulties was student talking time. Namely, if students did not talk for long enough, or their turns were too short, grading of student performance was considered difficult:

(R1) *Student 2 spoke less and I think he was the hardest to rate for it too.*

(R4) *Comparatively speaking student 4 is quite difficult for me to give a grade because she didn't have lots of turns.*

This was also a factor in non-verbal performance as well:

(R8) *…because you cannot really quantify how much time when they refer to notes will be regarded as whether they "usually" read from the notes, they "never" read from notes, etc.*

This factor has a particular impact on assessing discussion across the group, in that student talking time is related to impressions of the overall group dynamic as well.

(R10) *This group was a little more difficult to grade, there are two things. Student 5, her voice I guess wasn't projecting very well, It was hard to pick up a lot of her words, but student 2 and 3 sometimes 4, 2 and 3, their way to talk is very broken, they stop a lot, it caused the speech to be a little slower and more fragmented. I have to really listen to what they are saying*

(R6) *…because the bench mark is usually the best student in the group; that means the student that stands out, is the one who is the bench mark and that varies I think—considerably—from group to group.*

3.8 Reference to Rater

The coding scheme incorporated three potential challenges to the assessment in terms of whether opinions of general performance and raters" personal beliefs influence supposedly objective grading decisions, and whether teacher's practice might also influence their grade decisions. None of these considerations are reflected on the rubric, and so the usefulness of the rubric as a measurement of what is supposed to have been learned on the course may be called into question if raters can be shown to make grading decisions outside of the rubric's scope. The danger that raters" opinions pose to the validity of the test are often sourced in their subjective evaluations of "good" or "poor" performance, representing a (often unrelated) gut-instinct rather a fair representation of student performance on a given criterion over the allotted time. This can result in higher opinions of student performance, e.g.:

(R6) *When I watch what she is doing, it sounds good, she is thoughtful.*

(R7) *He's much more efficient, yes, that's much more efficient.*

But mostly results in lower opinions:

(R6)*…but I don't see anyone from this group seems to particularly understand the issue, its complexity, [there weren't] any sophisticated comments.*

(R11) *Student 3, he doesn't know what he is doing.*

(R2) *The level of discussion is pretty banal.*

(R10) *You know, she is interesting in the sense that she is fluent but she is not brilliant. Even if I'm just focusing on pronunciation, she is not brilliant…she is fluent, but there is something lacking there. Pretty good, but I've seen students better than her before.*

Higher or lower, these opinions stick with raters along the process of watching the video, and remain salient to them as they grade. A question may be raised here about whether raters are then objectively grading student performance against the criteria, or whether the raters, for whatever reason, first arrive at subjective appraisals of student performance, then attempt to make the criteria "fit" these subjective appraisals. Raters also often arrive at suspicious interpretations of student action during the assessment, which may or may not be supported in reality:

(R7) *It's not so much [about] reading the notes, but [whether they] have prepared a speech at the beginning. Like, student 1, I think clearly she had prepared a speech, and had used quite a lot of time.*

(R9) *So perhaps maybe she has some prepared speech ready, and that's why it doesn't really fit in to the discussion.*

Finally for this section, as the test raters are also teachers on the course, this appears to have an influence on some of their grading decisions:

> (R6) *My issue is trying to teach my students, if you don't organize the discussion, you can't really have a discussion. You've got to work out what the issue is, for the topic like this is ok, what are the roles of a journalist, let's see the role and we can start talking by each role, and then discuss each, the role of the journalist, that is high important so can we talk about that. I'm always looking out for that.*

> (R13) *That is something I pick up on, if students just say "I agree with you" and then just launch into their turn without really linking it to what came before. This is something that we see in the classes in the early mock speaking tests, and we try to teach our students not to do this, so if we do hear it in the test I think it is something we do pick up on.*

Thus, in terms of the construct validity of the assessment, there are appear to be a number of considerations that raters take into account when arriving at grading decisions that are not part of the assessment rubric, but are more in line with the subjective or experiential beliefs of the raters themselves.

3.9 Reference to Criteria

Overall, the themes related to difficulty with criterion validity were related to separating the individual sub-criteria both within and across the graded components (Stance, Interaction, Language), the grading of each sub-criteria within a graded component, the grading of criteria over time, and difficulties with the weighting of sub-criteria.

The single-most problematic criterion for raters was that of *"never dominate the discussion"*. This follows each of the statistical tests (ICC, CA, EFA) performed on the data, although the raters hardly mentioned the other problematic criterion *"never read from your notes"*. As for domination, raters were unsure as to how to define it, or how to measure it:

> (R13) *Whether that counts as domination or whether that's just counts as long talking time? The interpretation of domination according to what I know is about talking over people. That's something that they didn't do, there was no example where somebody is talking over someone else. I'm not sure, maybe we have different interpretations of domination*.

> (R3) *I would ask myself what "dominate" actually means, as I found it's quite difficult to grade student 5. I think overall student 5 has done a good job in discussion, but she is not concise enough, so I'm thinking whether she actually dominated the discussion. So, I lower her mark a little bit, but I think it's difficult to define "dominate"*.

Raters felt that some of the criteria within the graded components were hard to distinguish. The most problematic component in this regard was that of language, namely the relationship between fluency, grammar and comprehensibility:

> (R2) *the reason why is hard to grade fluency specifically is because when you are judging fluency I'm also taking into account their comprehensibility, vocabulary and grammar, I think it's kind of intertwined with the fluency, so it's hard to break up"*.

> (R13) *…but I would argue that having the grammar and vocabulary mistakes is like a separate consideration, but that bleeds into [whether you are] comprehensible or fluent.*

In terms of distinguishing criteria across the graded components, the results of the interviews concur with the majority of those of the factor analysis. In particular, the inclusion of *"Link contributions to what has been said"* in the interaction component was considered problematic, as raters felt unable to distinguish between this criterion and that of *"Critically respond to others"* in the stance component:

> (R12) *I think these two are very similar, "*critically respond to others" *and "*link contributions to what was said*" because sometimes we have to think about if someone has responded to someone else, doesn't that always mean it was linked to what was said? So, I think this was a bit hard to decide, one or two times I don't know where to put it [the grade]*.

> (R13) *Tying up the* "critical response to others" *and* "contribution had been linked to what has been said", *[I] struggle all the time to see the difference between those two.*

> (R14) *Generally if your contribution is linked to what it has been said then you [are] maybe critically responding to question others, but if you're doing that, you're also contributing, you're linking your contribution to what has been said before. So, those two for me seem to be tied"*.

Raters also felt that they were often double dipping for stance and grammar scores, which are supposed to be mutually exclusive on the rubric:

> (R14) *With student 5, the fluency isn't that good and I am picking up on that, and now I realise when I listen to the think aloud, that if students talk too slowly or hesitate, I stop listening for stance and start listening for fluency and grammar. But it seems like you can't listen for stance and fluency and grammar at the same time. […] Maybe we are not being fair to students who can't communicate that well because we are not listening to their stance but are too busy listening for fluency and grammar instead.*

A number of raters felt that time is an issue for grading of criteria, namely the tracking of performance over the 25 minutes for individual students on individual criteria. This was felt most keenly when grading stance:

> (R6) *It's very qualitative and you're trying to quantify something, you know for example one person can say something absolutely brilliant, an absolutely brilliant point, and after that [they] don't really saying anything particularly interesting. That one insight, one piece of critical insight or utterance, how much does it work? I don't know*.

> (R5) *Like fluency, I think that particular feature [is hard], because some of the students, he or she may be quite fluent in certain point of time, but towards the end of the assessment or discussion, the level of fluency is not very strong, so I found [that] very difficult.*

An additional issue was with weighting of the sub-criteria, a factor which is not explicitly dealt with in the rubric or in standardization meetings—the sub-criteria are (ideally) supposed to have equal weight. However, not all raters saw it this way:

> (R7) *There's a kind of a weighting to the sub-criteria as well, how much do you weight [sic] on that? I mean, how do you weigh it? Is critically responding to others worth 33%? How about explaining an academic concept? Which one is more cognitively difficult? It depends on the concept—it's a bit complicated.*

In summary, the qualitative data has provided some insight into the underlying causes of the quantitative findings, as well as uncovered additional validity concerns that lie outside of what could be determined quantitatively.

4. Discussion

For RQ1 and RQ2 (Consistency, consensus and measurement agreement, Criterion validity), the results of the validation study suggest that there are issues with the consistency, consensus and measure agreement in grading decisions across the two assessed performances, and that the configuration and wording of the assessment rubric may be a contributing factor for such variance. In terms of consensus, the ICC measures suggest that the rubric as a whole allows for a good level of consensus across raters, although two of the criteria related to interaction on the rubric result in a lack of rater consensus on grades, that of *never dominating the discussion* and *never looking at notes* during the assessment. In terms of consistency, a generally high level of consistency was achieved in the Cronbach's Alpha result, although the two criteria that caused problems for consensus (*domination*, *notes*) also served to lower the Cronbach's Alpha for the consistency of rater decision for the criteria as a whole. In terms of measurement, the EFA suggested that *never dominating the discussion* and *never looking at notes* were not part of the latent factor with high loadings for the other sub-criteria for stance and interaction, and stand alone in terms of the theoretical model. Problems defining "domination" alongside issues with quantifying behaviour over time serve as probable explanations in terms of the raters" interview feedback. An additional finding resulting from the EFA was that raters are unable to separate the grading of interaction from stance, and that the grading of *explaining academic concepts* is inseparable from that of grading for grammar. This results in raters double-dipping when marking stance/interaction and stance/grammar, despite the grading requiring that each main criteria be marked separately—an issue that was also a feature of the raters" interview feedback. These two main statistical findings, supported by qualitative data, cast doubt on the overall validity of the assessment.

In addition, the raters" interview feedback pointed out problems with rater reliability in terms of the affective and cognitive difficulties involved with assessing extended periods of interactional discourse. Factors including student talking time (or lack of it), determining exact (sub)grades from all available options, the group dynamic as a whole, subjective judgements of "good" performance outside of the criteria, and issues with the teacher as assessor were all features of the interview feedback, and constitute threats to validity that the statistical measurements were unable to capture.

In terms of RQ3 (how to improve the validity of the examination), based on the quantitative and qualitative findings of the validation process, a new sample rubric has been developed (Appendix B). The new rubric

incorporates the following main amendments:

1) As raters were unable to separate the grading of *stance* and *language* elements, such elements have been combined into a new main criteria titled "ideas". Here, the ability to explain academic concepts now requires accurate use of language. This criterion is worth 50% of the total grade.

2) The criteria for *interaction* have been modified to avoid confusion between "critically responding to others" (previously under *stance*) and "linking turns naturally", with these now both under the umbrella of *interaction*. A clearer, unambiguous definition of "domination" was required, now worded "talking over others". Sub-criteria that caused cognitive/affective difficulties such as "never read from notes" and "active listening" have been removed entirely. This new main criterion is worth 50% of the total grade.

3) Reference to scalar judgements of performance such as "always", "usually" etc. have been modified to detailed "can/cannot" type criteria, so as to avoid variance in grading decisions over extended periods of talk. This changes reduces the maximum number of individual grading decisions from fifty (ten sub-criteria across five letter grades) to ten. Studies such as Knoch (2009) note that detailed descriptors may result in substantially higher reliability measures, and the new criteria now reflect this consideration.

4) Reference to whether a student makes a full / substantial / partial or no contribution to the discussion is now present in the *interaction* criterion, addressing (part of) the issue with student talking time and grading. Students will now have to make more of a contribution in terms of talking time if they are to make a "full" contribution.

In terms of factors outside of the criteria, issues such as the group dynamic (in terms of students of mixed proficiency) and the "teacher as assessor" remain to be solved. The former may require more careful plotting of student performance over the entire course, so that low-fluency students who are unlikely to perform (comparatively) well against highly fluent students may be grouped and assessed together for the final assessment. Doing so may avoid raters" unfairly comparing high/low fluency students" individual performance and encourage low-fluency students to take centre stage where they might not ordinarily feel able to do so. However, this may provoke feelings of resentment among low-fluency learners if they feel unfairly labelled, or may cause resentment among high-fluency students if they feel that lower-fluency students may receive equivalent grades to them despite a linguistically poorer performance. In terms of the effect of the teacher-as-assessor dynamic on test validity, while there were few significant results of personal rater variables on grading decisions, teacher beliefs and differences in rater practice were significant features of the qualitative interviews. The new rubric is likely to go some way towards addressing these concerns in terms of carefully detailed "good" or "bad" performance, yet is likely that a standard procedure for reaching grading decisions has to be derived via increased moderation with pass/fail standards for rater variance that try (as much as possible) to ensure that rater prejudices do not impact their grading decisions.

Aside from changes to the rubric, course co-ordinators may wish to consider changes to the assessment standardisation and moderation process for group tutorial discussions. Changes to task conditions or rubrics must themselves be validated via revised training practice. In particular, greater consequences need to be faced by raters who fail standardisation activities, and more effort must be taken to lower the effect of rater background and beliefs on grading decisions. We also note that such changes may require more in the way of resources and challenges to staff morale if teaching staff are already working at their limits.

5. Conclusion

This validation study has outlined statistical and qualitative concerns relating to the consensus, consistency, measurement and validity of an academic group tutorial discussion assessment, leading to a new sample rubric. The study offers a cautionary lesson for language practitioners who create in-house rubrics without an accompanying validation process, and offers practical advice for those who are considering a switch to a group oral test format but are unsure how to implement an accompanying assessment practice.

We recognise that the sample size and scope in the present study are small, which may limit the generalisability of the study results, although we do consider the general findings of use to those considering implementing group tutorial discussion tasks as speaking assessments. In future research, we aim to capture the thought processes of raters as they view student interaction and make grading decisions in real time via think-aloud protocols, following Leung (2012), Davison (2004) and Ducasse & Brown (2009). Such an approach would allow us to determine specific moments in the overall discussion that triggered raters" attention, caused them to make positive or negative appraisals of student performance, see when and where raters reach agreement or disagreement, and show whether raters reach the same grading decisions based on different approaches, or reach

different grading decisions based on the same approach. Such a study is already underway.

References

Arkoudis, S., & O'Loughlin, K. (2004). Tensions between validity and outcomes: Teacher assessment of written work of recently arrived immigrant ESL students. *Language Testing*, *21*(3), 284-304. http://dx.doi.org/10.1191/0265532204lt285oa

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110. http://dx.doi.org/10.1191/0265532203lt245oa

Brindley, G. (1991). Defining language ability: the criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing*. Singapore: South East Asian Ministries of Education Organization.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46. http://dx.doi.org/10.1177/001316446002000104

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory* (Vol. 6277). New York: Holt, Rinehart and Winston.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51. http://dx.doi.org/10.1177/026553229000700104

Davison, C. (1999). Missing the mark: the problem with benchmarking ESL in Australian schools. *Prospect*, *14,* 66-76.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, *21*(3), 305-334. http://dx.doi.org/10.1191/0265532204lt286oa

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423-443. http://dx.doi.org/10.1177/0265532209104669

East, M. (2006). The impact of bilingual dictionaries on lexical sophistication and lexical accuracy in tests of L2 writing proficiency: A quantitative analysis. *Assessing Writing, 11*(3), 179-197. http://dx.doi.org/10.1016/j.asw.2006.11.001

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York, NY: Wiley.

Galaczi, E. (2004). *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English.* Unpublished PhD dissertation, Teachers College, Columbia University, New York.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference.* 11.0 update (4th ed.). Boston: Allyn & Bacon

Gliem, J., & Gliem, R. (2003). *Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales.* Paper presented at the 2003 Midwest Research to Practice Conference in Adult, Continuing and Community Education, the Ohio State University, Columbus, Ohio.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. Dechert & G. Raupach (Eds.), *Interlingual Processes* (pp. 229-244). Tubingen: Gunter Narr.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*(1), 5-16. http://dx.doi.org/10.1016/S1075-2935(02)00029-6

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse studies, 7*(2), 173-192. http://dx.doi.org/10.1177/1461445605050365

Hyland, K. (2016). Writing with attitude: Conveying a stance in academic texts. In E. Hinkel (Ed.), *Teaching English Grammar to Speakers of Other Languages* (pp. 246-265). London: Routledge.

Johnson, M. (2001). *The art of non-conversation: A re-examination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Kirkpatrick, A. (2007). *World Englishes paperback with audio CD: Implications for international communication and English language teaching*. Cambridge: Cambridge University Press.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26* (20), 275-304. http://dx.doi.org/10.1177/0265532208101008

Kohn, A. (2006). The trouble with rubrics. *English Journal, 95*(4), 12-15. http://dx.doi.org/10.2307/30047080

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies, 48*(6), 661-671. http://dx.doi.org/10.1016/j.ijnurstu.2011.01.016

Krashen, S. D. (1987). *Principles and Practices in Second Language Acquisition.* New York: Prentice-Hall.

Landers, R. N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower, 2:e143518.81744*.

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System, 20*, 373-386. http://dx.doi.org/10.1016/0346-251X(92)90047-7

Leung, C. (2012b). Qualitative research in language assessment. *Encyclopedia of Applied Linguistics*. Retrieved from http://onlinelibrary.wiley.com.eproxy2.lib.hku.hk/doi/10.1002/9781405198431.wbeal0979/pdf

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Mesa Press.

Long, M. (1985). Input and Second Language Acquisition Theory. In S. M. Gass & C. G. Madden (Eds.), *Input and Second Language Acquisition* (pp. 377-393). Rowley, M.A.: Newbury House.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*(3), 246-276. http://dx.doi.org/10.1191/0265532202lt230oa

Pica, T. (1987). Second-language acquisition, social interaction, and the classroom. *Applied linguistics, 8*(1), 3-21. http://dx.doi.org/10.1093/applin/8.1.3

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing writing, 15*(1), 18-39. http://dx.doi.org/10.1016/j.asw.2010.01.003

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390. http://dx.doi.org/10.1177/0265532207077205

Saxton, E., Belanger, S., & Becker, W. (2012). The Critical Thinking Analytic Rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assessing Writing, 17*(4), 251-270. http://dx.doi.org/10.1016/j.asw.2012.07.002

Schmidt, R. (1992). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13*, 206-226. http://dx.doi.org/10.1017/S0267190500002476

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1-19.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Los Angeles: Sage. http://dx.doi.org/10.4135/9781412995627.d5

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied linguistics, 16*(3), 371-391. http://dx.doi.org/10.1093/applin/16.3.371

Taylor, L. (2001). The paired speaking test format: Recent studies. *UCLES Research Notes, 6*. Retrieved from http://www.cambridgeesol.org/rs_notes/rs_nts6.pdf

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508. http://dx.doi.org/10.2307/3586922

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23,* 411-440. http://dx.doi.org/10.1191/0265532206lt336oa

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511732997

Notes

Note 1. The rules of thumb for Cronbach's Alpha range from " > .9 – Excellent, > .8 – Good, > .7 –

Acceptable, > .6 – Questionable, > .5 – Poor, and < .5 – Unacceptable" (George & Mallery, 2003, p. 231)

Note 2. A coded segment can be coded under more than one part of the scheme at the same time.

Appendix A

Speaking Test Assessment Criteria *(2015-16 Revised)*

	A+, A, A-	B+, B, B-	C+, C, C-	D+, D	F
Ability to explain academic concepts and argue for a stance supported by sources **40% of grade**	You can **always** clearly explain academic concepts. You are **always** able to argue for a critical stance with the support of valid academic sources where appropriate. You show an **excellent** ability to critically respond to / question other students' stance.	You can **almost always** clearly explain academic concepts. You are **usually** able to argue for a critical stance with the support of valid academic sources where appropriate. You show a **good** ability to critically respond to / question other students' stance.	You are **usually able to** explain academic concepts but **sometimes not** clearly. While you can **usually** argue for a stance, it is **not very** detailed or supported by any academic sources and is **usually** simplistic rather than critical. You show a **limited** ability to critically respond to / question other students' stance.	There is **only some evidence** of an ability to explain academic concepts and these are **usually** unclear. There is **only some evidence** of an ability to argue for a stance and when you do, it is **almost always** simplistic rather than critical and not supported by any academic sources. You show a **limited** ability to critically respond to / question other students' stance and when you do attempt to, the meaning is unclear. You are **mostly silent** throughout the discussion.	What you say is **almost always** unclear. You are **unable** to express a stance. You **never** critically respond to other students' stance. You have **no notesheet /** You have **plagiarized** your notesheet from another student. You **never** use any sources.
Ability to interact with others **30% of grade**	You **never** dominate the discussion. You **never** read from your notes when expressing your stance. Your contributions to the discussion are **always** naturally linked to what has been said before. You **always** use active listening skills (nodding, eye contact etc.) when appropriate.	You **never** dominate the discussion. You **never** read from your notes when expressing your stance. Your contributions to the discussion are **almost always** naturally linked to what has been said before. You **almost always** use active listening skills (nodding, eye contact etc.) when appropriate.	You dominate the discussion in **one or two places**. You **sometimes** read from your notes when expressing your stance. Your contributions to the discussion are **usually** naturally linked to what has been said before. You **usually** use active listening skills (nodding, eye contact etc.) when appropriate.	You **often** dominate the discussion. You **often** read from your notes when expressing your stance. Your contributions to the discussion are **only sometimes** naturally linked to what has been said before. You **only sometimes** use active listening skills (nodding, eye contact etc.) when appropriate.	Your interaction skills are **too limited** to be able to successfully take an active role in the tutorial discussion.
Ability to communicate comprehensibly and fluently **30% of grade**	You are **always** comprehensible. Mistakes with grammar / vocabulary are **infrequent** and **never** interfere with understanding. You are **always** fluent.	You are **nearly always** comprehensible. Mistakes with grammar / vocabulary are **infrequent** and **rarely** interfere with understanding. You are **usually** fluent.	You are **generally** comprehensible. Mistakes with grammar / vocabulary **occur throughout** but **rarely** interfere with understanding. You are **generally** fluent.	You are **only sometimes** comprehensible. Mistakes with grammar / vocabulary **occur throughout** and interfere with understanding **in multiple places.** You are **only sometimes** fluent.	Your spoken language causes **repeated and sustained** strain on the listener.

Appendix B

Sample Draft New Assessment Rubric

	A+, A, A-	B+, B, B-	C+, C, C-	D+, D, D-	Fail
Ideas (50%)	Can fully explain and critically evaluate academic concepts with excellent support from the literature and with no difficulty in understanding across the entire discussion	Can explain academic concepts well with useful support from the literature, although occasionally may lack criticality or have minor citation or language problems	Explains academic concepts, although often superficially, with little useful support from the literature, or has major citation / language problems at times	Weak presentation of academic concepts with very little / no support from the literature, or often struggles to be understood due to language problems	Candidate cannot explain academic concepts, does not use any literature, or frequently cannot be understood due to language issues. Has plagiarised or not prepared a notesheet.
Interaction (50%)	Makes a full contribution to the entire discussion by linking turns naturally to what was said, questioning and critically responding to others' stance while defending their own. Does not talk over others.	Makes a substantial contribution to the discussion but occasionally does not link turns to what was said or fails to respond to others' stance. May occasionally talk over others	Makes a contribution to the discussion but does not often link their turn to what was said or only occasionally responds to others' stance. May talk over others more than necessary.	Contributes little to the discussion, either because they say too little, or do not link their turns well to what was said, preferring their own conversation rather than responding to others' stance. May frequently interrupt others mid-speech.	Contributes nothing to the discussion, or does not consider the others in the discussion at any time.